Code Submissions

1. **Github Repository**: [Link]
2. **requirements.txt**

```
streamlit
scikit-learn
pandas
numpy
matplotlib
seaborn
xgboost
```

README.md

Breast Cancer Prediction & Analysis Pipeline

a. Problem Statement

The objective of this assignment is to implement an end-to-end Machine Learning classification pipeline. This involves training multiple classification models on a chosen dataset to predict a target variable, evaluating their performance using various metrics, and deploying the best-performing models via an interactive Streamlit web application. The goal is to demonstrate proficiency in the entire ML workflow: data selection, preprocessing, modeling, evaluation, and deployment.

b. Dataset Description

Dataset Name: Breast Cancer Wisconsin (Diagnostic) Dataset
Source: sklearn.datasets (originally from UCI Machine Learning Repository)
Description: Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.
Target: Diagnosis (M = malignant, B = benign)
Features: 30 numeric features (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension).
Instances: 569 (Meets requirement of >= 500)
Feature Count: 30 (Meets requirement of >= 12)

c. Models Used & Comparison Table

The following 6 classification models were implemented and evaluated:

1. Logistic Regression
2. Decision Tree Classifier
3. K-Nearest Neighbor (kNN) Classifier
4. Naive Bayes Classifier (Gaussian)
5. Random Forest Classifier (Ensemble)
6. XGBoost Classifier (Ensemble)

Evaluation Metrics Comparison

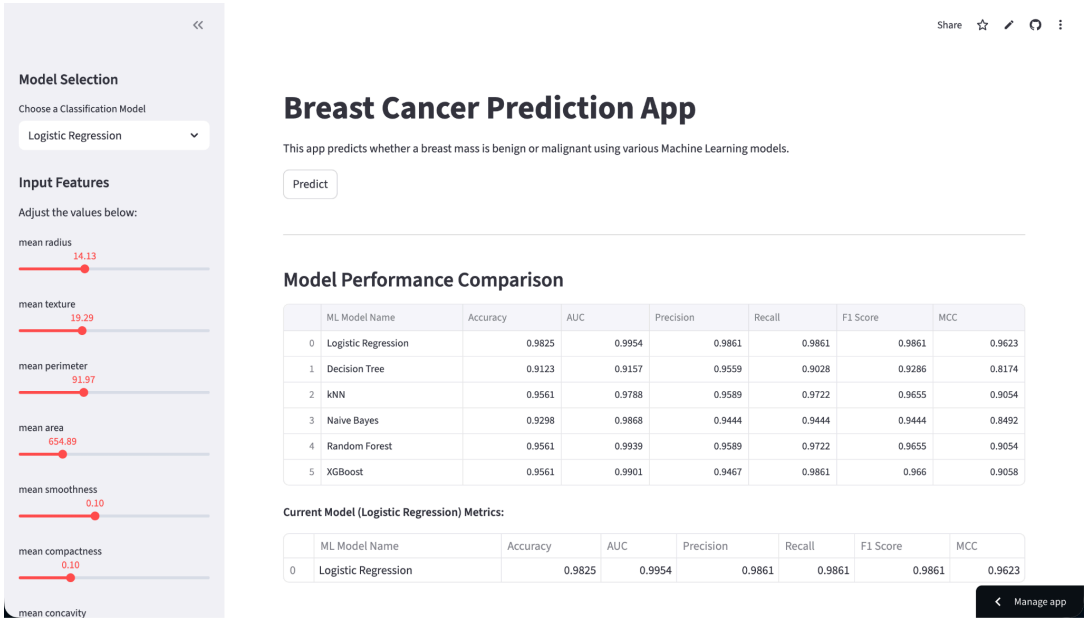
ML Model Name	Accuracy	AUC Score	Precision	Recall	F1 Score	MCC Score
Logistic Regression	0.9825	0.9954	0.9861	0.9861	0.9861	0.9623
Decision Tree	0.9123	0.9157	0.9559	0.9028	0.9286	0.8174
kNN	0.9561	0.9788	0.9589	0.9722	0.9655	0.9054
Naive Bayes	0.9298	0.9868	0.9444	0.9444	0.9444	0.8492
Random Forest	0.9561	0.9939	0.9589	0.9722	0.9655	0.9054
XGBoost	0.9561	0.9901	0.9467	0.9861	0.9660	0.9058

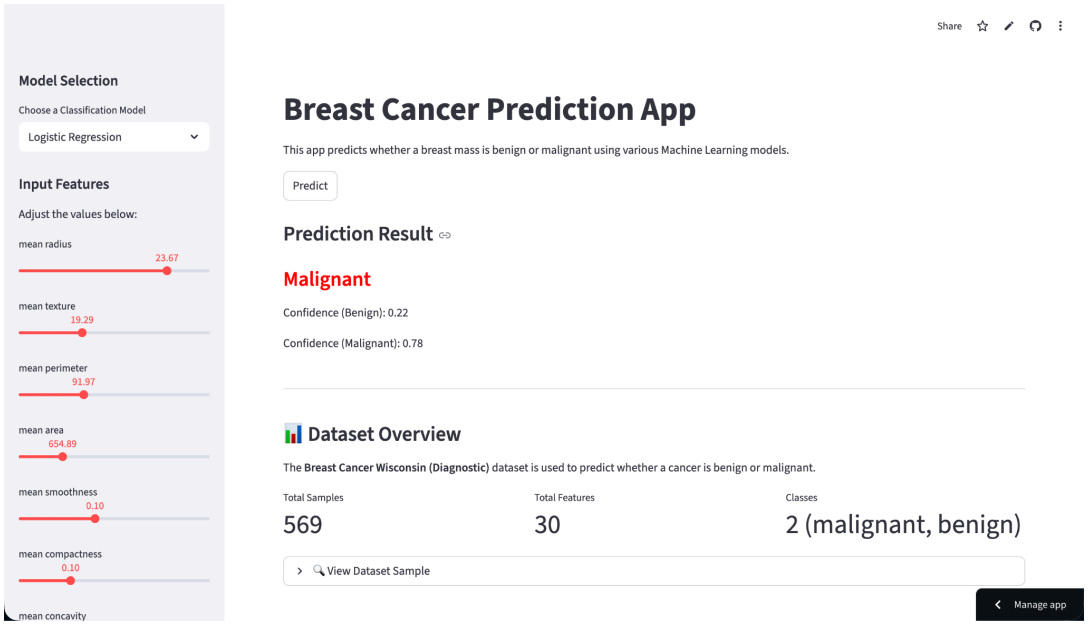
Observations about Model Performance

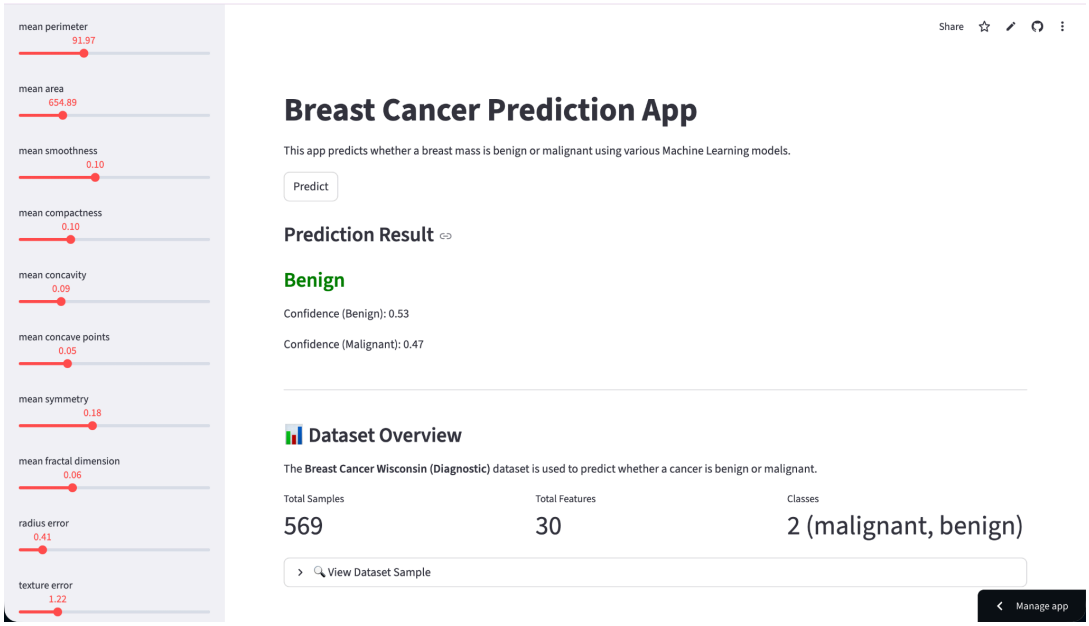
1. **Logistic Regression**: Performed exceptionally well, achieving the highest Accuracy (98.25%) and F1 Score (98.61%). This suggests the dataset is linearly separable to a high degree.
2. **Ensemble Models (Random Forest & XGBoost)**: Both performed robustly with identical Accuracy (95.61%). They handle non-linear relationships well but were slightly outperformed by the simpler Logistic Regression on this test set.
3. **Decision Tree**: Had the lowest accuracy (91.23%) among the models, likely due to overfitting on the training data compared to the ensemble methods which mitigate this.
4. **Naive Bayes**: Performed reasonably well (93%) given its strong independence assumptions, showing that the features are likely independent enough for this model to be effective.
5. **kNN**: Achieved competitive results (95.61%), similar to the ensemble models, indicating that local neighborhood structures are preserving class information well.
6. **Overall**: All models achieved >90% accuracy, making them suitable for this task. Logistic Regression is the recommended model for this specific dataset and split due to its simplicity and superior performance.

Streamlitt App

1. **Application: [Link]**
2. **Application Screenshots:**

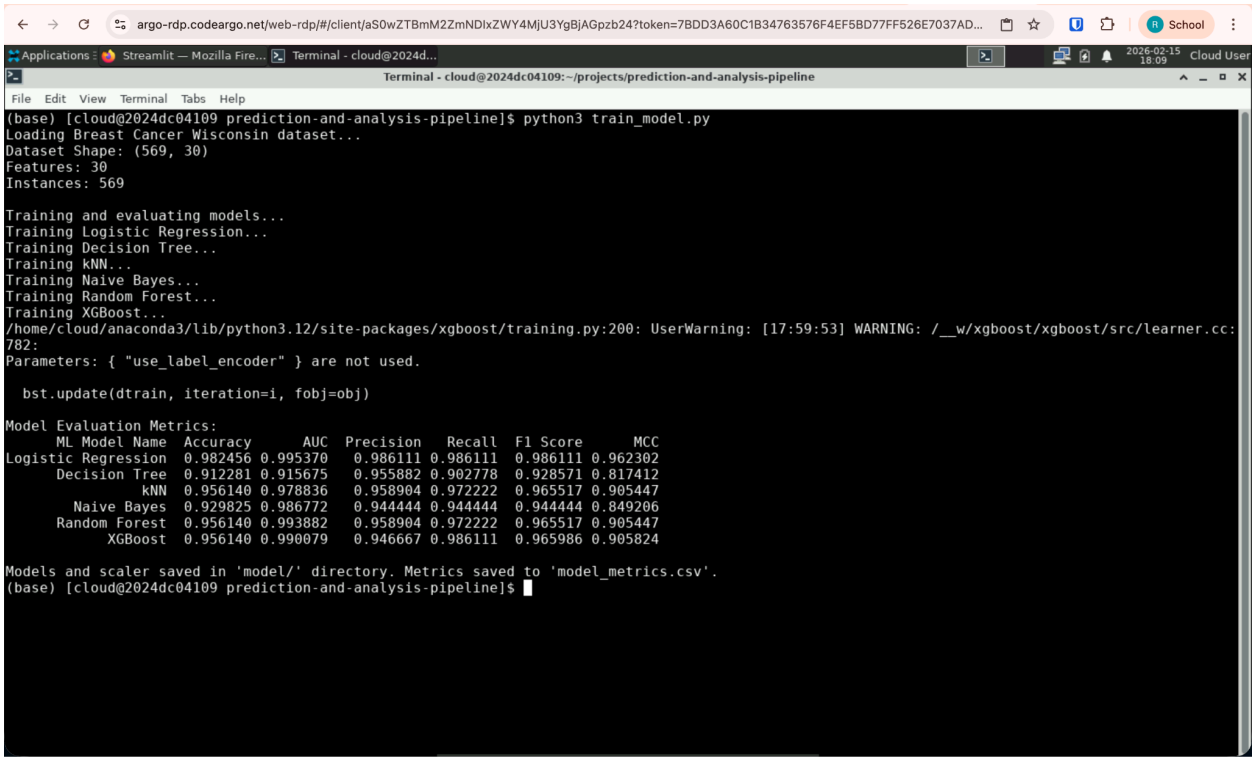






Lab Screenshots

1. Terminal Screenshots after training model



2. Application Running on Localhost